Field-based species identification of closely-related plants using real-time nanopore sequencing

Parker, Joe; Helmstetter, Andrew J.; Davey, Dion; Wilkinson, Tim; Papadopulos, Alexander S. T.

Scientific Reports

DOI:
10.1038/s41598-017-08461-5

Published: 01/08/2017

Publisher's PDF, also known as Version of record

SCIENTIFIC REPORTS

OPEN

Field-based species identification of closely-related plants using real-time nanopore sequencing

Received: 3 July 2017
Accepted: 12 July 2017
Published online: 21 August 2017

Joe Parker[1], Andrew J. Helmstetter[1], Dion Devey[1], Tim Wilkinson[1] & Alexander S. T. Papadopulos[1,2]

Advances in DNA sequencing and informatics have revolutionised biology over the past four decades, but technological limitations have left many applications unexplored. Recently, portable, real-time, nanopore sequencing (RTnS) has become available. This offers opportunities to rapidly collect and analyse genomic data anywhere. However, generation of datasets from large, complex genomes has been constrained to laboratories. The portability and long DNA sequences of RTnS offer great potential for field-based species identification, but the feasibility and accuracy of these technologies for this purpose have not been assessed. Here, we show that a field-based RTnS analysis of closely-related plant species (*Arabidopsis* spp.) has many advantages over laboratory-based high-throughput sequencing (HTS) methods for species level identification and phylogenomics. Samples were collected and sequenced in a single day by RTnS using a portable, "al fresco" laboratory. Our analyses demonstrate that correctly identifying unknown reads from matches to a reference database with RTnS reads enables rapid and confident species identification. Individually annotated RTnS reads can be used to infer the evolutionary relationships of *A. thaliana*. Furthermore, hybrid genome assembly with RTnS and HTS reads substantially improved upon a genome assembled from HTS reads alone. Field-based RTnS makes real-time, rapid specimen identification and genome wide analyses possible.

DNA sequencing used to be a slow undertaking, but the past decade has seen an explosion in HTS methods[1,2]. DNA barcoding (i.e., the use of a few, short DNA sequences to identify organisms) has benefited from this sequencing revolution[3–5], but has never become fully portable[6–8]. Samples must be returned to a laboratory for testing and the discrimination of closely related species using few genes can be problematic due to evolutionary phenomena (e.g. lineage sorting, shared polymorphism and hybridisation)[2]. While typical barcoding approaches have been effective for generic level identification, accuracy is much more limited at the species level[3,9] and concerns remain[10]. Species delimitation using limited sequencing information has also been problematic and is thought to heavily underestimate species diversity[3,11]. Consequently, increasingly elaborate analytical methods have been spawned to mitigate the inherent limitations of short sequences[9,12]. The Oxford Nanopore Technologies® MinION® is one of a new generation of RTnS DNA sequencers that is small enough to be portable for fieldwork and produces data within minutes[13–17]. These properties suggest species identification could be conducted using genome scale data generated at the point of sample collection. Furthermore, the large number of long reads generated[13] may provide more accurate species-level identification than current approaches. This application offers great potential for conservation, agriculture, environmental biology, evolutionary biology and combating wildlife crime. However, this potentially exciting combination of methods has not yet been rigorously tested in the field for eukaryotic genomes.

Our experiment was designed to determine whether DNA reads produced entirely in the field could accurately identify and distinguish samples from closely-related species (*A. thaliana* (L.) Heynh. and *A. lyrata* (L.) O'Kane & Al-Shehbaz). Recent analyses have shown that gene flow has been common and shared polymorphisms are abundant between the morphologically distinct species in *Arabidopsis*. Indeed, the two study species share >20,000 synonymous SNPs[18], making this a good stress test of genome scale RTnS sequencing for species discrimination.

[1]Jodrell Laboratory, Royal Botanic Gardens, Kew, Richmond, Surrey, UK, TW9 3AB. [2]Molecular Ecology and Fisheries Genetics Laboratory, Environment Centre Wales, School of Biological Sciences, Bangor University, Bangor, UK, LL57 2UW. Correspondence and requests for materials should be addressed to J.P. (email: joe.parker@kew.org) or A.S.T.P. (email: a.papadopulos@kew.org)

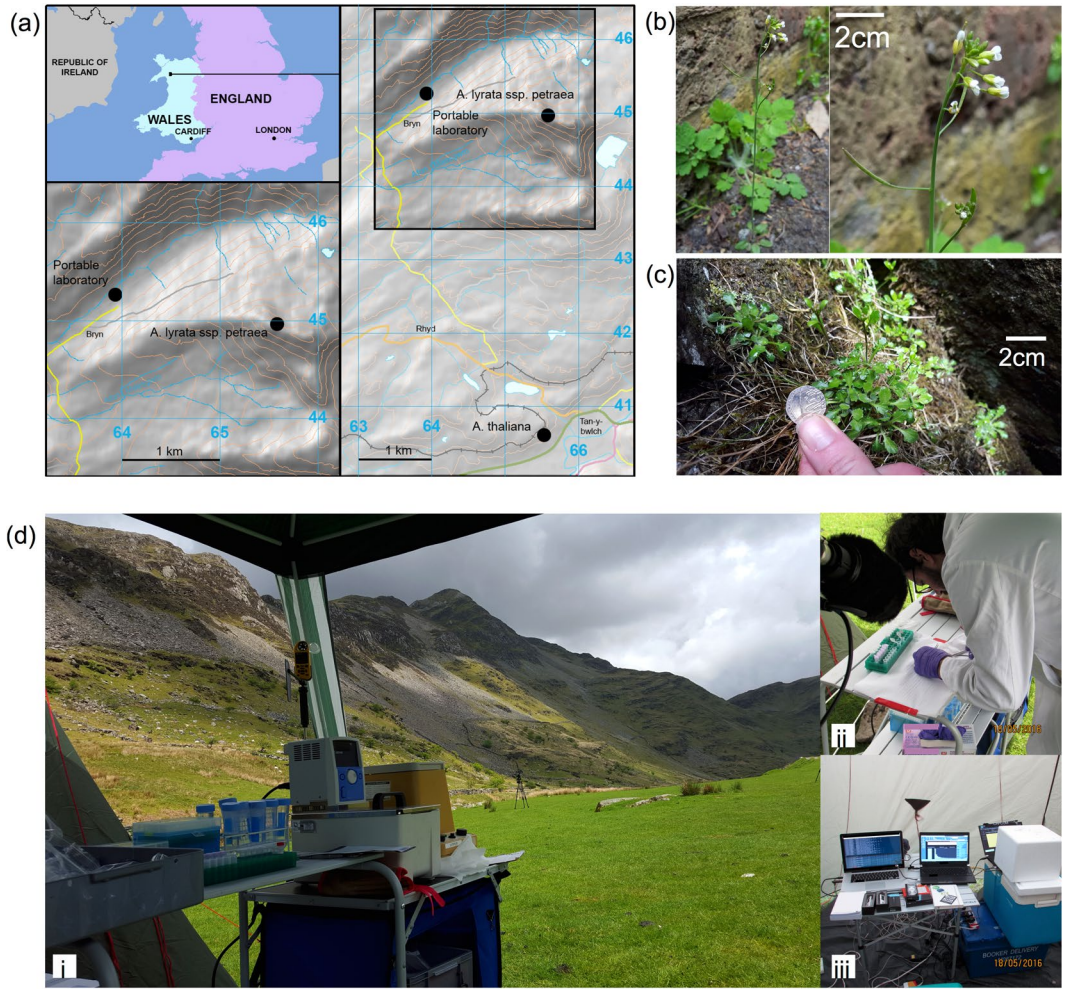

Figure 1. Logistics and scope of field-based sequencing. (**a**) Location of sample collection and extraction, sequencing and analyses in the Snowdonia National Park, Wales. Maps were created using ESRI ArcGIS Desktop v10.5. Source of elevation data: U.S. Geological Survey, Shuttle Radar Topography Mission 1 Arc-Second Global[46]. (**b**) *Arabidopsis thaliana*. (**c**) *A. lyrata* ssp. *petraea*. (**d**) The portable field laboratory used for the research. Ambient temperatures varied between 7–16 °C with peak humidity >80%. A portable generator was used to supply electrical power.

Results and Discussion

The first goal was to extract and sequence shotgun genomic data from higher plant species in the field using RTnS technology in sufficient quantity for downstream analyses within hours of the collection of plant tissue (Extended Fig. 1). On consecutive days, tissue was collected from three specimens each of *A. thaliana* and *A. lyrata subsp. petraea* (Fig. 1b,c) in Snowdonia National Park, and prepared, sequenced and analysed outdoors in the Croesor Valley (Fig. 1a). Only basic laboratory equipment was used for DNA extraction and MinION sequencing-library preparation; we did not use a PCR machine (Fig. 1d; Extended Data Table 1). One specimen of each species was sequenced with both R7.3 and R9 MinION chemistries. For *A. thaliana*, the RTnS experiment generated 97 k reads with a total yield of 204.6Mbp over fewer than 16 h of sequencing (see Extended Data Table 2). Data generation was slower for *A. lyrata*, over ~90 h sequencing (including three days of sequencing at RBG Kew following a 16 h drive), 26 k reads were generated with a total yield of 62.2Mbp. At the time, a limited implementation of local basecalling was available for the R7.3 data only. Of 1,813 locally basecalled reads, 281 had successful BLAST matches to the reference databases with a correct to incorrect species ID ratio of 223:30. The same samples were subsequently sequenced using HTS short read technology (Illumina MiSeq™, paired-end, 300 bp; Supplementary Note 1). Mapping reads to available reference genomes for the *A. thaliana* (TAIR10 release[19]) and two *A. lyrata* assemblies[20,21] indicates approximate RTnS coverage of 2.0x, 0.3x, and 0.3x for *A. thaliana*, *A. lyrata*, and *A. lyrata ssp. petraea*, respectively; and 19.5x, 11.9x and 12.0x respectively for HTS reads (Extended Data Tables 2 and 3, Supplementary Note 2). These results demonstrate that the entire process (from sample collection thorough to genome scale sequencing) is now feasible for eukaryotic species within a few hours in field conditions. As the technology develops, run yields are expected to improve and implementation of sample indexing will allow many samples to be run on a single flow cell.

As expected given the developmental stages of the technologies, the quality and yield of field sequenced RTnS data was lower than the HTS data (Extended Data Tables 2 and 4). *Arabidopsis thaliana* RTnS reads could be aligned to approx. 50% of the reference genome (53Mbp) with an average error rate of 20.9%. Indels and mismatches were present in similar proportions. The *A. lyrata* RTnS data were more problematic with significantly poorer mapping to the two *A. lyrata* assemblies, whereas, the HTS data performed relatively well. For the limited number of alignable RTnS reads, error rates were slightly higher than for *A. thaliana* (22.5% and 23.5%). The poorer RTnS results for *A. lyrata* may be a consequence of temperature-related reagent degradation in the field or due to unknown contaminants in the DNA extraction that inhibited library preparation and/or RTnS sequencing. Despite the smaller yield and lower accuracy of the RTnS compared to HTS data, the RTnS reads were up to four orders of magnitude longer than the HTS reads and we predicted they would be useful for species identification, hybrid genome assembly and phylogenomics.

To explore the utility of these data for species identification, the statistical performance of field-sequenced (RTnS) and lab-sequenced (HTS) read data was assessed. Datasets for each species were compared to two databases via BLASTN, retaining single best-hits: one database contained the *A. thaliana* reference genome and the second was composed of the two draft *A. lyrata* genomes combined. Reads which matched a single database were counted as positive matches for that species. The majority of matching reads hit both databases, which is expected given the close evolutionary relationships of the species. In these cases, positive identifications were determined based on four metrics; (a) the longest alignment length, (b) the highest % sequence identities and (c) the largest number of sequence identities (d) the lowest *E*-value (Extended Data Table 4). Test statistics for each of these metrics were calculated as the difference of scores (length, % identities, or *E*-value) between 'correct' and 'incorrect' database matches. The performance of these difference statistics for binary classification was assessed by investigating the true and false positive rates (by reference to the known sample species) across a range of threshold difference values (Fig. 2a–d and Extended Data Figs 2–4). For both short- and long-read data at thresholds greater than 100 bp, the differences in total alignment lengths (ΔL_T) or number of identities (ΔL_I) are superior to e-value or % identity biases (Fig. 2a–d). Furthermore, at larger thresholds (i.e., more conservative tests), RTnS reads retained more accuracy in true- and false-positive discrimination than HTS data. This proves that whole genome shotgun RTnS is a powerful method for species identification. We posit that the extremely long length of the observed 'true positive' alignments compared with an inherent length ceiling on false-positive alignments in a typical BLASTn search is largely responsible for this property.

To evaluate the speed with which species identification can be carried out, we performed *post hoc* analyses by subsampling the RTnS *A. thaliana* dataset. This simulated the rate of improvement in species assignment confidence over a short RTnS run. We classified hits among the subsampled reads based on (i) ΔL_I over a range of threshold values (ii) mean ΔL_I and (iii) aggregate ΔL_I (Fig. 3). This demonstrates that a high degree of confidence can be assigned to species identifications over the timescales needed to generate this much data (i.e., $<$one hour) and that variation in the accuracy of identifications quickly stabilises above 1000 reads. Aggregate ΔL_I values rapidly exclude zero (no signal) or negative (incorrect assignment) values, making this simple and rapidly-calculated statistic particularly useful for species identification. In a multispecies context, the slopes of several such log-accumulation curves could be readily compared, for example (see Supplementary Discussion).

The *Arabidopsis* reference genomes used here were highly contiguous (N50 $>$ 20Mbp), but the majority of reference genomes (both published and unpublished) are considerably more fragmented. To understand the effect of reference genome contiguity on species identification with our method, we repeated these analyses after simulating increasingly fragmented reference genomes from the published assemblies. As expected, alignment length and number of identities biases did decrease at lower N50 values, but not precipitously. Even the poorest assembly simulated (N50 ~ 1,000 bp) is sufficient to identify the species confidently (see Extended Data Fig. 5). This suggests that rapid generation of low coverage genome assemblies can provide adequate information for species identification with field-sequenced long read data.

Assembly of large and complicated eukaryotic genomes with RTnS data alone would require a greater volume of data than available here[7,22–24]. Field extracted samples are unlikely to be of similar purity to those obtained with more sophisticated laboratory-based methods, leading to lower yields. As expected, *de novo* assembly of our RTnS data performed poorly, likely due to insufficient coverage. However, these data do have potential for hybrid genome assembly approaches. We assembled the HTS data *de novo* using ABYSS[25] and produced a hybrid assembly with both RTnS and HTS datasets using HybridSPAdes[26]. The hybrid assembly was an improvement over the HTS-only assembly (see Extended Data Table 5) with fewer contigs, a total assembly length closer to the reference (1190Mbp), N50 and longest contig statistics both increasing substantially and estimated completeness (CEGMA[27]) of coding loci increased to ~99%. These results suggest that relatively small quantities of long and short reads can produce useful genome assemblies when analysed together, an important secondary benefit of field–sequenced data. The length of typical RTnS reads is similar to that of genomic coding sequences (1–10 kb)[13]. This raises the possibility of extracting useful phylogenetic signal from such data, despite the relatively high error rates of individual reads. We annotated individual raw *A. thaliana* reads directly, without genome assembly, which recovered over 2,000 coding loci from the data sequenced in the first three hours (Fig. 2e). These predicted gene sequences were combined with a published dataset spanning 852 orthologous, single-copy genes[28], downsampled to 6 representative taxa. Of our gene models, 207 were present in the Wickett *et al.*[28] dataset and the best 56 matches were used for phylogenomic analysis (see Supplementary Methods for details). The resulting phylogenetic trees (Fig. 2f and Extended Data Fig. 6) are consistent with the established intergeneric relationships[28]. Although the taxonomic scale used here for phylogenomics is coarse it highlights an additional benefit to rapid, in-the-field sequencing for evolutionary research.

This experiment is the first to demonstrate field-based sequencing of higher plant species. When directly compared to lab-based HTS, our experiment highlights key discriminatory metrics for highly accurate species identifications using portable RTnS sequencing. Few approaches can boast this level of discriminatory power

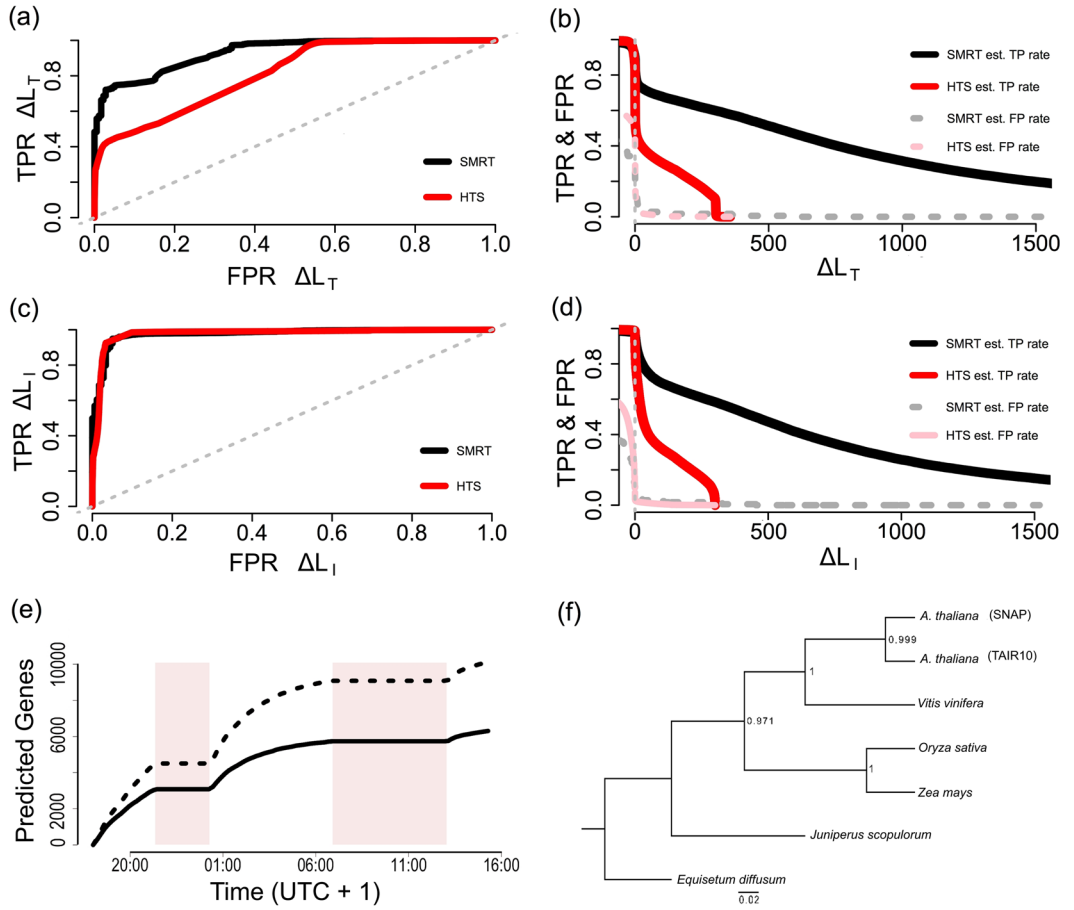

Figure 2. Sample identification and phylogenomics using field-sequenced RTnS data. (**a**–**d**) Orthogonal species identification using BLASTN difference statistics: HTS data (red) and RTnS (black) matched to reference databases via BLASTN. (**a**,**c**) Receiver operating characteristic (ROC; estimated false-positive rate vs. estimated true positive rate) and (**b**,**d**) estimated true- (solid lines) and false-positive (dashed lines) rates. (**a**,**b**) ΔL_T statistic; (**c**,**d**) ΔL_I statistic. (**e**) Accumulation curves for *ab initio* gene models predicted directly from individual *A. thaliana* reads over time. Count of unique TAIR10 genes (solid line) and total number of gene models (dashed line). Shaded boxes represent periods where the MinION devices were halted while the laboratory was dismantled and moved. (**f**) phylogenetic tree inferred under the multispecies coalescent from RTnS reads.

and none of these have the same degree of portability[2,3]. The data produced for identification is also useful for genome assembly. Finally, entire coding sequences can be recovered from single reads and incorporated into evolutionary analyses. Clearly, data generated with the goal of accurate species identification has much broader usefulness for genomic and evolutionary research. Few technical barriers remain to prevent the adoption of portable RTnS by non-specialists, or even keen amateurs and schoolchildren. As these tools mature, and the number of users expands, portable RTnS sequencing can revolutionise the way in which researchers and practitioners can approach ecological, evolutionary and conservation questions.

Methods

Study site and sample collection. On consecutive days, tissue was collected from three specimens each of *A. thaliana* and *A. lyrata subsp. petraea* in Snowdonia National Park and sequenced and analysed in a tent. *A. lyrata* was collected from the summit of Moelwyn Mawr (52.985168° N, 4.003754° W; OL17 65554500; SH6558244971) and *Arabidopsis thaliana* was collected at Plâs Tan-y-Bwlch (52.945976° N 4.002730° W; OL18 65604060; SH6552940610). Representative voucher specimens of each species are deposited at RBG, Kew. DNA extractions, library preparation and DNA sequencing with the MinION technology were all conducted using portable laboratory equipment in the Croesor valley on the lower slopes of Moelwyn Mawr immediately following sample collection (52.987463°N 4.028517° W; OL17 63904530; SH6392745273). Laboratory reagents were stored in passively-cooled polystyrene boxes with internal temperatures monitored using an Arduino Uno. Only basic laboratory equipment was used (including two MinION sequencers and three laptops; see Extended Data Table 1).

DNA extraction. The standard Qiagen DNeasy plant mini prep kit was used to extract genomic DNA from *Arabidopsis spp.* with the exception that the two batches were pooled at the DNeasy mini spin column step to maximise the DNA yield.

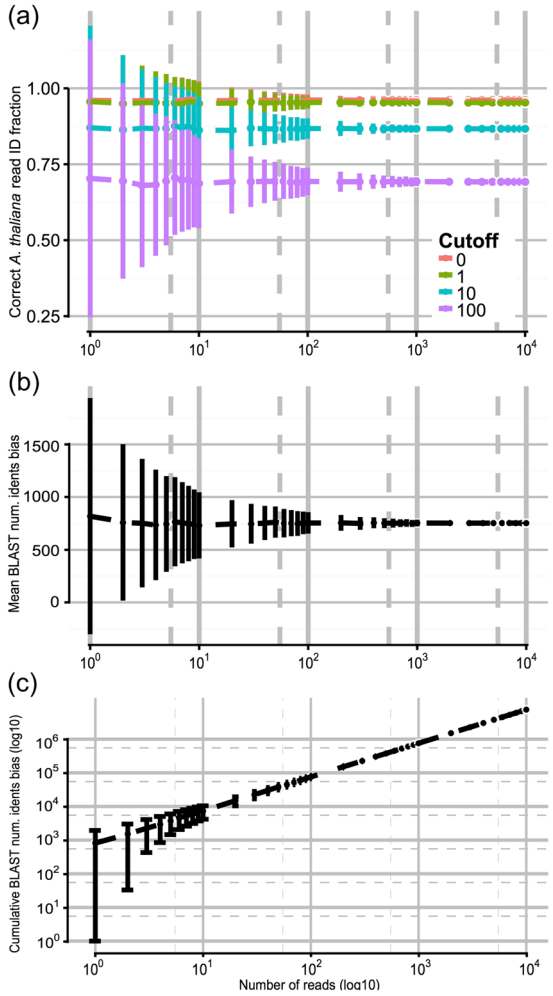

Figure 3. Simulated accumulation curves for rapid species identification by DNA sequencing. 34k pairwise BLASTN hits of *A. thaliana* RTnS reads were subsampled without replacement to simulate an incremental accumulation of data (10^4 reads; 10^3 replicates). For each read the total identities bias (ΔL_I) is the number of identities with the *A. thaliana* reference minus the number of identities with the *A. lyrata* reference. (**a**) the proportion of *A. thaliana* reads correctly identified on a per-read basis, classified as *A. thaliana* where $\Delta L_I >$ threshold cutoff (0, 1, 10 or 100). (**b**) Mean ΔL_I in the simulated dataset rapidly stabilises on the population mean (+754 bp, e.g. an average matching read alignment to *A. thaliana* is 754 bp longer than to *A. lyrata*). (**c**) Cumulative aggregate Δ_{LI}; negative or zero ΔL_I can rapidly be excluded. Typical data throughput rates exceed 10^4 reads per hour of sequencing.

DNA library preparation and sequencing. An R7.3 and R9 1D MinION library preparation were performed for each species according to the manufacturer's instructions using a developer access programme version of the commercially available Nanopore RAD-001 library kit (Oxford Nanopore Technologies). No PCR machine was used. Lambda phage DNA was added to *A. thaliana* R9 library for quality control. For *A. thaliana*, the MinION experiment generated 96,845 1D reads with a total yield of 204.6Mbp over fewer than 16 h of sequencing. Data generation was slower for *A. lyrata*, possibly due to temperature-related reagent degradation or unknown contaminants in the DNA extraction. Over ~90 h sequencing, 25,839 1D reads were generated with a total yield of 62.2Mbp; this included three days of sequencing at RBG Kew following a 16 h drive, during which reagents and flowcell were stored sub-optimally (near room-temperature). BLASTN 2.4.0[29] was used to remove 5,130 reads with identity to phage lambda. Data are given in Extended Data Table 2. The following week in a laboratory, NEBNext Ultra II sequencing libraries were prepared for four field-extracted samples (two individuals from each species) and sequenced on an Illumina MiSeq (300 bp, paired end). In total, 11.3Gbp and 37.8 M reads were generated (each ~8 M reads and 2Gbp; see Supplementary Note 1).

Field offline basecalling and bioinformatics in real-time. Offline basecalling using nanocall 0.6.13[30] was applied to the R7.3 data as no offline R9 basecaller was available at the time. Basecalled reads were compared to the reference genomes of *A. thaliana* (TAIR10 release) and *A. lyrata subsp petraea* (1.0 release). In total, 119 reads were processed in real-time with six reads making significant hits by BLASTN that scored correctly:

incorrectly for species ID in a 2:1 ratio. After the sequencer had been halted a larger dataset of 1,813 reads gave 281 hits, with correct: incorrect: tied identifications in a 223:30:28 ratio.

Accuracy and mapping rates of short- and long-read data. Both lab-sequenced NGS reads (trimmed with Trimmomatic[31]) and untrimmed, field-sequenced RTnS reads were aligned to the appropriate reference genomes using the BWAv0.7.12-r1039[32] and LASTv581[33], to estimate depth of coverage and nominal error rate in mapped regions (see Supplementary Note 2). For all *A. thaliana* datasets (short and long-read), average mapped read depths were approximately equal to the gross coverage. MinION reads could be aligned to 53Mbp of the reference genome with LAST (approx. 50% of the total genome length). The nominal average error rate in these alignments was 20.9%). For both MinION and MiSeq datasets, mapping and alignment to the *A. lyrata* and *A. lyrata ssp. petraea* assemblies was more problematic. For alignable MinION reads, error rates were slightly higher than for *A. thaliana* at 22.5% and 23.5%, estimated against *A. lyrata* and *A. lyrata ssp. petraea* assemblies, respectively. We note that these assemblies are poorer quality than the *A. thaliana* TAIR10 release; total genome lengths differ (206Mbp and 202Mbp,) and contiguity is relatively poor in both (695 and 281,536 scaffolds).

Determination of true- and false-positive detection rates, sensitivity, and specificity. Each of the four datasets (HTS and RTnS, for each species) was matched against two custom databases (the *A. thaliana* reference genome and the two draft *A. lyrata* genomes combined) separately with BLASTN, retaining only the best hit for each query. Queries matching only a single database were counted as positive matches for that species (Extended Data Table 4). Non-matching reads were treated as negative results (Supplementary Methods). Queries matching both databases were defined as positives based on: a) longest alignment length (L_T); b) highest % sequence identities, c) longest alignment length counting only identities (L_I), or c) lowest E-value. Test statistics for each of these metrics were simply calculated as the difference of scores (length (ΔL_T), % identities, identities (ΔL_I), or E-value) between 'true' and 'false' hits. The statistical performance of these statistics (true- and false-positive rates, and accuracy) in putative analyses under varying threshold values were calculated and visualized using the ROCR package in R[34]. The high proportion of reads with significant hits to both species is expected given the close evolutionary relationships of the species. Analyses to determine the best statistics to discriminate between species using reads which aligned to both databases strongly indicated that difference in alignment lengths between the best discriminator, shown in Fig. 2a–d and Extended Data Figs 2, 3 and 4. Overall these show that the difference in alignment length is a powerful indicator for both short- and long-read data at any threshold $\geq$ ~100 bp. Furthermore, and surprisingly, at this and more conservative (greater difference) threshold, long-read field-sequenced reads had substantially more accuracy in true- and false-positive discrimination than short-read data. This suggests that this method provides a powerful means of species identification and we posit that the extremely long length of 'true positive' alignments compared with the natural length ceiling on false-positive alignments is largely responsible for this property.

Accumulation curves for simulated identification. 33,806 pairwise BLASTN hits obtained above in identification against *A. thaliana* and *A. lyrata* genomic reference databases were subsampled without replacement to simulate incremental accumulation of BLASTN hit data during progress of a hypothetical sequencing experiment producing 10,000 reads produced in total. 1,000 replicates were used to calculate means and variances for data accumulation in 0.1 log-increments from $r = 1$ read to 10^4 reads total. For each read, ΔL_I, 'number of identities bias', was calculated as the difference (number of identities in *A. thaliana* alignment – number of identities in *A. lyrata* alignment). Each read was assigned to *A. thaliana* or not if it ΔL_I exceeded a given threshold, repeated at four possible values, $L_{threshold} = \{0, 1, 10, 100\}$. Mean and aggregate (total) ΔL_I values were also calculated for each replicate over the progress of the simulated data collection. Results are shown in Fig. 3.

Genome assembly fragmentation for simulated identification. The *A. thaliana* (TAIR10; N50 = 23.5Mbp) and *A. lyrata* (1.0; N50 = 24.5Mbp) reference genomes were fragmented *in silico* without replacement. Fragment lengths were picked from a uniform distribution parameterised to produce simulated N50 lengths of 10^3, 10^4, 10^5 and 10^6 bp, with three replicates per N50 length. The resulting simulated assemblies were used to create separate nucleotide BLAST databases and the read ID procedure repeated as above using the simulated databases. Full script and details are available at GitHub: http://github.com/lonelyjoeparker/real-time-phylogenomics/wales-analyses/in-silico-genome-skimming.

***De novo* genome assembly.** Short-read HTS data was assembled *de novo* using ABYSS v1.9.0[35]. A hybrid assembly with both HTS and RTnS datasets was performed with HybridSPAdes v3.5.0[36]. Assemblies were completed for *A. thaliana* (sample AT2a) and *A. lyrata* (sample AL1a). Assembly statistics were calculated in Quast v4.3[37]. Completeness of the final hybrid assemblies was assessed using CEGMA v2.5[38]. Results of *de novo* genome assemblies are given in Extended Data Table 5. Analyses of genome contiguity and correctness and conserved coding loci completeness indicated that assembly of HTS data performed as expected (20x coverage produced ~25,000 contigs covering approximately 82% of the reference genome at an N50 of 7,853 bp). By contrast, the hybrid assembly of *A. thaliana* illumina MiSeq and Oxford Nanopore MinION data significantly improved on the MiSeq-only assembly: 24,999 contigs reduced to 10,644; total assembly length increased to close to the length of the reference genome (119.0Mbp) with nearly 89% mappable; N50 and longest contig statistics both improved (N50 7,853 → 48,730 bp) indicating better contiguity from the addition of long reads. Completeness of coding loci as estimated by CEGMA (Extended Data Table 5) greatly increased to ~99%. Long reads did not compromise the accuracy of high-coverage short-read data; basewise error rates were not significantly worse.

Direct gene annotation of single unprocessed field-sequenced reads. The length of typical individual RTnS reads is of similar magnitude to genomic coding sequences. Consequently, useful phylogenomic information could potentially be obtained by annotating reads directly, without a computationally expensive genome assembly step. Raw, unprocessed *A. thaliana* reads were individually annotated directly without assembly via SNAP[39]. To verify which gene predictions were genuine, the DNA sequences (and 1 kb flanking regions, where available) were matched to available *A. thaliana* (TAIR10) genes with default parameters. BLAST hits were further pruned based on quality (based on 1st-quartile quality scores: alignments length bias $\Delta L_T \geq +570$ bp/% identities bias $\geq +78.68/E$-value bias ≥ 0), reducing the number of hits from 18,098 to 10,615. Sample read alignments and details of SNAP output BLAST score summary statistics are given in Supplementary Table 1 and encounter curves-through-time are shown in Fig. 2e.

Phylogenomics of raw-read-annotated *A. thaliana* genes. Predicted *A. thaliana* gene sequences were combined with a published phylogenomic dataset spanning 852 orthologous, single-copy genes in plants and algae[28], downsampled to 6 representative taxa for speed: *Equisetum diffusum, Juniperus scopulorum, Oryza sativa, Zea mays, Vitis vinifera* and *A. thaliana*. Our putative gene models were assigned identity based on reciprocal best-hit BLASTN matching with the *A. thaliana* sequences in these alignments, yielding 207 matches, of which the top 56 were used for phylogenomic analysis (Supplementary Table 1), only 18 having no missing taxa in the Wickett *et al.*[28] dataset. Alignments were refined using MUSCLE v3.8.31[40] and trimmed with a 50% missing-data filter (using trimAL v1.4rev15[41]) then used to infer species trees in two ways: (i) single gene phylogenies inferred separately (using RAxML v7.2.8[42]) under the GTRCAT substitution model with 10 discrete starting trees then combined into a summary tree using TreeAnnotator v.1.7.4[43]; (ii) a species tree inferred directly from the data under the multispecies coalescent[44], implemented in *BEAST v2.4.4[45] (with adequate MCMC performance confirmed using Tracer v1.5). A maximum clade credibility (MCC) tree was produced using TreeAnnotator v.1.7.4. Phylogenies inferred by orthodox (RAxML) and multispecies coalescent (*BEAST) methods are shown in Extended Data Figure 5 and agreed with each other and the established phylogeny presented in Wickett *et al.*[28].

Data availability. Basecalled read data for Illumina and Oxford Nanopore sequencing runs are available via the EBI ENA at PRJEB22018. Analyses and custom scripts used are deposited in GitHub at https://github.com/lonelyjoeparker/realtime-phylogenomics/tree/master/wales_analyses.

References

1. Hajibabaei, M., Baird, D. J., Fahner, N. A., Beiko, R. & Golding, G. B. A new way to contemplate Darwin's tangled bank: how DNA barcodes are reconnecting biodiversity science and biomonitoring. *Philos. Trans. R. Soc. London B Biol. Sci.* **371**, 20150330 (2016).
2. Mallo, D. & Posada, D. Multilocus inference of species trees and DNA barcoding. *Philos. Trans. R. Soc. London B* **371**, 20150335 (2016).
3. CBOL Plant Working Group *et al.* A DNA barcode for land plants. *Proc. Natl. Acad. Sci. USA* **106**, 12794–7 (2009).
4. Hollingsworth, P. M., Li, D.-Z., van der Bank, M. & Twyford, A. D. Telling plant species apart with DNA: from barcodes to genomes. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150338 (2016).
5. Hebert, P. D. N., Hollingsworth, P. M., Hajibabaei, M. & Hebert, P. D. N. From writing to reading the encyclopedia of life. *Philos. Trans. R. Soc. London B Biol. Sci.* **371**, 1–9 (2016).
6. Schmidt, K. *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J. Antimicrob. Chemother.* dkw397, doi:10.1093/jac/dkw397 (2016).
7. Datema, E. *et al.* The megabase-sized fungal genome of Rhizoctonia solani assembled from nanopore reads only. *bioRxiv*, doi:10.1101/084772 (2016).
8. Erlich, Y. A vision for ubiquitous sequencing. *Genome Res.* **25**, 1411–1416 (2015).
9. Little, D. P. DNA barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLoS One* **6** (2011).
10. Collins, R. A. & Cruickshank, R. H. The seven deadly sins of DNA barcoding. *Mol. Ecol. Resour.* **13**, 969–975 (2013).
11. Tang, C. Q. *et al.* The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. *Proc. Natl. Acad. Sci.* **109**, 16208–16212 (2012).
12. Zhang, A. B. *et al.* A fuzzy-set-theory-based approach to analyse species membership in DNA barcoding. *Mol. Ecol.* **21**, 1848–1863 (2012).
13. Laver, T. *et al.* Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **3**, 1–8 (2015).
14. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. USA* **93**, 13770–3 (1996).
15. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–32 (2016).
16. Faria, N. R. *et al.* Mobile real-time surveillance of Zika virus in Brazil. *Genome Med.* **8**, 97 (2016).
17. Edwards, A., Debbonaire, A. R., Sattler, B., Mur, L. A. & Hodson, A. J. Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N. *bioRxiv* 73965, doi:10.1101/073965 (2016).
18. Novikova, P. Y. *et al.* Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).
19. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).
20. Hu, T. T. *et al.* The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476–81 (2011).
21. Akama, S., Shimizu-Inatsugi, R., Shimizu, K. K. & Sese, J. Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. *Nucleic Acids Res.* **42** (2014).
22. Goodwin, S., Gurtowski, J., Ethe-sayers, S., Deshpande, P. & Michael, C. Oxford Nanopore Sequencing and de novo Assembly of a Eukaryotic Genome. (2015).
23. Mikheyev, A. S. & Tin, M. M. Y. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.* **14**, 1097–1102 (2014).
24. Jansen, H. J. *et al.* Rapid de novo assembly of the European eel genome from nanopore sequencing reads. *bioRxiv*, doi:10.1101/101907 (2017).
25. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E. & Jones, S. J. M. ABySS: A parallel assembler for short read sequence data ABySS: A parallel assembler for short read sequence data. 1117–1123, doi:10.1101/gr.089532.108 (2009).
26. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. HybridSPAdes: An algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).

27. Parra, G., Bradnam, K. & Korf, I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
28. Wickett, N. J. *et al*. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci.* **111**, E4859–E4868 (2014).
29. C. Coulouris, G. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2008).
30. David, M., Dursi, L. J., Yao, D., Boutros, P. B. & Simpson, J. T. Nanocall: An Open Source Basecaller for Oxford Nanopore Sequencing Data. *Bioinformatics* **33**(1), 49–55, doi:10.1093/bioinformatics/btw569 (2017).
31. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30**(15), 2114–2120 (2014).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009).
33. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**(3), 487–93 (2011).
34. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: Visualizing classifier performance in R. *Bioinformatics* **21**(20), 3940–3941 (2005).
35. Simpson, J. T. *et al*. ABySS: a parallel assembler for short read sequence data. *Genome research* **19**(6), 1117–1123 (2009).
36. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**(7), 1009–15, doi:10.1093/bioinformatics/btv688 (2016).
37. Gureyvich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**(8), 1072–1075 (2013).
38. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**(9), 1061–1067 (2007).
39. Korf, I. Gene finding in novel Genomes. *BMC Bioinformatics* **5**, 59 (2004).
40. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5), 1792–97 (2004).
41. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Salvador Capella-Gutierrez; Jose M. Silla-Martinez; Toni Gabaldon. *Bioinformatics* **25**, 1972–1973 (2009).
42. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9), 1312–1313 (2014).
43. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology And Evolution* **29**, 1969–1973 (2012).
44. Heled, J. & Drummond, A. J. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27**(3), 570–580 (2010).
45. Bouckaert, R. *et al*. BEAST 2: A software platform for bayesian evolutionary analysis. *PLoS Computational Biology* **10**(4), e1003537 (2014).
46. Farr, T. *et al*. The shuttle radar topography mission. *Rev. Geophys.* **45**, 1–33 (2007).

Acknowledgements

This work was funded by a Pilot Study Grant to JDP and a Howard Lloyd Davies legacy grant to ASTP. JDP was also supported by funding from the Calleva Foundation Phylogenomic Research Programme and the Sackler Trust. The authors also thank The Botanical Society of Britain & Ireland. Natural Resources Wales, Robyn Cowan, and Patricia and David Brandwood for assistance.

Author Contributions

A.S.T.P. and J.D.P. conceived the study and obtained funding. A.S.T.P., D.D. and J.D.P. designed and conducted fieldwork. A.S.T.P. designed and conducted field-based labwork with input from J.D.P., A.J.H. and D.D. A.J.H. conducted lab-based sequencing. J.D.P. conducted bioinformatics and phylogenomic analyses with contributions from A.J.H. and A.S.T.P. T.W. generated the cartographic data and images. A.S.T.P. and J.D.P. prepared the manuscript with contributions from A.H., D.D. and T.W.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-08461-5

Competing Interests: Oxford Nanopore Technologies contributed MinION sequencing reagents and flowcells for this research. JDP and ASTP received travel remuneration and free tickets to present an early version of this work at a conference (London Calling 2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.